PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Use of Simulation to Assess Electronic Health Record Safety in the Intensive Care Unit-A Pilot Study
AUTHORS	gold, Jeffrey; March, Christopher; Steiger, David; Scholl, Gretchen; Mohan, Vishnu; Hersh, William

VERSION 1 - REVIEW

REVIEWER	Gong, Michelle Montefiore Medical Center, Critical Care Medicine
REVIEW RETURNED	02-Feb-2013

THE STUDY	March CA and his coauthors present the results from a very interesting simulation study aimed at determining the proficiency by which medical trainees recognize and report errors of commission and omissions in a multi-day course of a simulated EHR ICU patient with sepsis and ARDS. The authors designed a study of simulation in an area where simulation is not usually utilized but perhaps, should be used more frequently. The authors tested 38 interns, residents and fellows and found that they detected an average of 41% of the 14 possible medical errors or conditions. They claimed that the amount of errors correlated with training level but the error rate is so high as to suggest an inherent problem with EHR interface in its ability to communicate information effectively. Indeed, they found the number of errors detected correlated with the number of different screens used and certain high impact screens. This article makes several notable contributions to the current discussion on patient safety and physician training and education in the ICU. Most notably, it highlights the high number of errors that can occur with the vast volume of data generated in the ICU and in EHR. However, the manuscript would be even stronger if the authors can address several issues that needs better clarification. 1. It is not clear why the authors choose to test multiple days of data. In simulation of uncommon events like difficult airways or cardiac resuscitation, a test in a control environment with a test patient is very effective. However, in simulations that involve common routine interactions and workflow activities, it is important that the simulation closely mimic the usual workflow and routine so as to more accurately represent the environment. As such it may be more realistic to simulate morning rounds or sign outs between residents during change of shift. Usually, in these cases, patient data over the last 24 hours rather than over multiple days are reviewed and presented during ICU rounds or sign out. This is especially true with the increasing shift work of residents in the ICU given work hour restrictions. Thus recognition of trends over time may be even lower in reality. The authors should consider discussion of how the simulation may and may not mimic usual workflow in their ICU and possible implication for true error recognition rate.

	2. The authors assert that much of the errors missed is due to suboptimal interfaces or training on use of EHR to abstract data and avoid cognitive overload. To provide proper context for this interpretation, the investigators should describe what training is given to trainees on ICU EHR and how much prior experience in the ICU and most recent rotation in the ICU. Most EHR, EPIC included, have different screens for ICU patients than for floor patients so many of the trainees may not have had much experience with the ICU interfaces. In addition, it is important to show screen shots of the EHR interfaces especially the "high impact" screens such as the Synopsis and MD index screens in Figure 5. This is needed for readers to better understand why performance scores may different with these two screens. Other specific comments: Other specific comments by section are below: Title: The title suggests that this study was designed to test EHR safety, yet the objective in the abstract is "to test the efficiency". These are two different topics. Please clarify the main objective. Key Messages: There is a typo in key message 3 Introduction: This is very well-written and a nice review. However, it should be shortened and the parts most germane to this study should be highlighted. Methods: There is a typo "The case was designed with the central theme of a determining" You do not include the "H" in FASTHUG. Was this not counted as an error? Was FASTHUG uniformly taught in the ICU? Was it part of a daily checklist? In Table 1, only glucose control and sedation in FASTHUG were included as possible error. The investigators may be more accurate by removing FASTHUG from the methods and just focusing on glucose control and sedation management in the text. More information should be provided about the experience of the subjects. Had the interns had previous ICU experience? Were these junior or senior fellows? The simulation was tested with 2 senior fellows. However, the fact that they could complete it in 10 minutes does not indicate that an intern can complete the same task in 10 minutes. The residents were given a signout before examining the medical record. Was this a standardized signout?
RESULTS & CONCLUSIONS	1. While their finding that the participants detected an average of 41% of the errors is concerning, the investigators' study design may actually be overly optimistic. For example, the fact that the participants were observed during their data collection and were aware that they would have to present the case later undoubtedly creates a significant Hawthorne effect. The participants may be more vigilant about their data collection because they are aware of being directly observed and tested. In addition, because discussion of EHR use has been integrated into weekly M & M conference, it is likely that the participants, especially the fellows, are more cognizant of the errors that commonly occur with EHR than trainees at other institutions with EHR. While this would bias the results towards improved recognition of errors, this still needs to be discussed further in the discussion precisely because the low rate of error

	recognition is so alarming. It is important to demonstrate that potential biases from study design would actually bias the results towards the null suggesting that the true error recognition rate may be even lower in actual clinical practice. 2.The authors should clarify that the inability of trainees to recognize errors or deteriorating medical condition does not necessarily imply that poor outcomes will result. The errors may be picked up by the pharmacist, the attending or nurse or it may be detected later. Indeed, while “massive amount of information” with data overload is one problem with standard EHRs, error rate and its implied consequences may be exaggerated in this simulation because one person was responsible for finding every error rather than a team of clinicians. Other specific comments about the results, discussion: Results: The EHR is just one tool a clinician uses. Some things such as the sedation level of the patient may be easier to assess by a physical exam rather than by searching the EHR. Other things may have been found out through other methods like communication with other health professionals. Discussion: In citation 22 the subjects were not asked question related to “whether the patient was bleeding”. They were informed that the patient was bleeding and then asked questions like is there adequate IV access, is there a type and screen active. Would it be better to perform simulation training with end-users before purchasing an EHR? Tables: Some of the items should be detailed and quantified further (increase in WBCs, new fever, recognition of fluid balance)
REPORTING & ETHICS	I would be willing to write an editorial on this paper if published.

REVIEWER	Michael W. Smith, PhD Health Science Specialist/Human Factors Engineer Houston VA HSR&D Center of Excellence Baylor College of Medicine USA I have no competing interests.
REVIEW RETURNED	07-Feb-2013

GENERAL COMMENTS	This is a very nice study which touches on the important issues of EHR design, data overload, and detection of medical errors. The use of a rich, complex case, and the face-valid method of eliciting physicians’ assessments, are valuable contributions. I have some suggestions for how to make the paper stronger: **GOALS** One goal appears to be about using simulation to assess how providers using an EHR can detect evidence of medical errors in the record (p2, Article Focus). So your outcomes are detection rates, plus a secondary outcome of patterns of navigation in the EHR.

On p 17 line 56, you mention that “the goal of the simulation is to test the system under high-stress/dangerous situations” (and I assume “test” means testing the user/EHR system on the task of problem detection). You mention this it as a mitigation of a limitation, but I would argue that if your goal is to look at patterns of problem detection, then having a case with lots of problems may be appropriate.

On p8, the goal is described as development of simulated ICU patient encounter in EHR*, as part of an effort to train better use of EHR so providers can better detect errors and manage data overload. (*I would describe it more as a simulated handoff or rounding task involving a rich multi-day record of a complex case).

You also introduce training on p7, and you mention the lack of any training intervention results as a limitation of the study (p2).

Your design also includes comparisons across clinical training levels (fellows vs. residents vs. interns).

The study demonstrates an assessment method, rather than a training intervention. If you are going to spend time on the role of simulation in training, I think you should be more clear and explicit about how this study fits in.
Using simulators in training, and using simulators for assessment of training (or for assessment of user interface designs) are different things.

The difference between the clinical levels is very interesting, and demonstrates a type of validity (that your protocol can detect differences between these groups). However, you are not explicit about how the comparison across clinical levels fits in with your goals.

FYI: One way differences in expertise are used in instructional systems R&D is to study the patterns of use that the more successful practitioners do, and use that to inform the content of training for less experienced people. [See: Crandall B, Klein G, Hoffman R. *Working Minds: A Practitioner's Guide to Cognitive Task Analysis (Bradford Books)*. The MIT Press; 2006.]. What did the fellows do that (presumably) helped them cope with the information overload better?
Related to this: Is there any more you can say about how some people used a larger number of screens in the same 10 minutes? Navigation techniques are one way people cope adapt to information overload [Watts-Perotti J, Woods D. How Experienced Users Avoid Getting Lost in Large Display Networks. *International Journal of Human-Computer Interaction*. 1999;11(4):269-299.]

FIDELITY
All simulations represent a sub-set of reality. The question of fidelity is really about which sub-set of reality, and which mode of representation, is relevant to the research question. See:

- Rudolph JW, Simon R, Raemer DB. Which reality matters? Questions on the path to high engagement in healthcare simulation. *Simulation in Healthcare*. 2007;2(3):161-163.; and

- McCurdy M, Connors C, Pyrzak G et al. Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. ACM; 2006:1233-1242.

Given that your research question is about problem detection via the electronic record, the fact that the test took place in the same physical environment as is normally used may not be especially relevant to your research question. One simulation constraint that may be relevant is that the physician received a written summary of the patient's history, rather than a warm handoff. What questions would an incoming physician ask the outgoing physician about the patient? Would communication with other staff also affect problem detection?

Note that the types of cases (and even the modes of representation) useful for training are not necessarily the same as those for assessment.

ANALYSIS & RESULTS

I'd like to see more on which medical errors are considered serious and which are not. Relatedly, you use categories in Table 1 (changes in patient condition, medication errors, failure to adhere to best practice), but these categories are not used in presentation of results.

Can you report anything about the participants' familiarity with EPIC, or EHRs in general?

I'm not sure if having both figure 2 and figure 3 is necessary.
With figure 2, % correct (instead of N correct vs. N incorrect) would be more consistent with some of the other figures.

Figure 5 should use the more informative terms used in text.

MISC
Missing word p 5 line 44
Typo p2 Line 34

VERSION 1 – AUTHOR RESPONSE

Reviewer #1

It is not clear why the authors choose to test multiple days of data. In simulation of uncommon events like difficult airways or cardiac resuscitation, a test in a control environment with a test patient is very effective. However, in simulations that involve common routine interactions and workflow activities, it is important that the simulation closely mimic the usual workflow and routine so as to more accurately represent the environment. As such it may be more realistic to simulate morning rounds or sign outs between residents during change of shift. Usually, in these cases, patient data over the last 24 hours rather than over multiple days are reviewed and presented during ICU rounds or sign out. This is especially true with the increasing shift work of residents in the ICU given work hour restrictions. Thus recognition of trends over time may be even lower in reality. The authors should consider discussion of how the simulation may and may not mimic usual workflow in their ICU and possible implication for true error recognition rate.

We appreciate the reviewers comments and we agree that data are usually only presented for the last 24hrs during a typical signout, the interpretation of the significance of those data can only be taken into context of the baseline status of the patient. As an example, in our cases, even though changes in hemodynamics occur over 24hrs, their absolute values are above what would be considered a “danger threshold” (eg MAP of 60mm Hg or SBP of 90mm Hg) but represent a significant drop (30%) reduction from the patient’s established baseline. This trend can best be appreciated in a contextual fashion when data beyond the prior 24 hours is evaluated by the subject.

2. The authors assert that much of the errors missed is due to suboptimal interfaces or training on use of EHR to abstract data and avoid cognitive overload. To provide proper context for this interpretation, the investigators should describe what training is given to trainees on ICU EHR and how much prior experience in the ICU and most recent rotation in the ICU. Most EHR, EPIC included, have different screens for ICU patients than for floor patients so many of the trainees may not have had much experience with the ICU interfaces. In addition, it is important to show screen shots of the EHR interfaces especially the “high impact” screens such as the Synopsis and MD index screens in Figure 5. This is needed for readers to better understand why performance scores may different with these two screens.

This is an excellent point. In our implementation of EPIC, there are no specific screens for the ICU as compared to inpatient rotations. Further, while the interns had variable experience with the system, all of the fellows and residents had over a year working with the system including multiple months in an ICU environment. This may explain some of the slight reduction in performance as well between interns and the other two groups and the discussions section has been modified to reflect this. In terms of training, all users received identical institution -mandated training at the beginning of training. This has also been clarified in the methods section. Finally, while we would love to show screen shots of the high impact screens, typically EHR vendors consider using their product screenshot images as an intellectual property / business concern and usually do not grant permission to share these or any other screen shots from the system. We felt it was important to disseminate the results of this pilot study, and will certainly engage with the EHR vendor in future to obtain their permission to include product screenshots.

Other specific comments:

Other specific comments by section are below:
Title:
The title suggests that this study was designed to test EHR safety, yet the objective in the abstract is “to test the efficiency”. These are two different topics. Please clarify the main objective.
We apologize for this confusion and this has been clarified

Key Messages:
There is a typo in key message 3
This has been corrected.

Introduction:
his is very well-written and a nice review. However, it should be shortened and the parts most germane to this study should be highlighted.

Methods:

There is a typo “The case was designed with the central theme of a determining”
This has been corrected

You do not include the “H” in FASTHUG. Was this not counted as an error? Was FASTHUG uniformly taught in the ICU? Was it part of a daily checklist? In Table 1, only glucose control and sedation in FASTHUG were included as possible error. The investigators may be more accurate by removing FASTHUG from the methods and just focusing on glucose control and sedation management in the text.

We apologize for this exclusion, HEAD of Bed elevation is part of the FASTHUG. The FASTHUG was part of daily rounds and our daily checklist. However, it should be noted that as we designed this study (with the goal of developing additional cases) one class of errors to include in all cases are deviation of daily standards of care. We have used the FASTHUG as defining potential candidates for this category based on its widespread international acceptance and our institutional use of the acronym. Therefore, an error was not included to account for every aspect of the FASTHUG, rather, our definition of deviations from daily best practice are items covered by the FASTHUG. We have clarified this in the methods section.

More information should be provided about the experience of the subjects. Had the interns had previous ICU experience? Were these junior or senior fellows?
This has been added to the method section.

The simulation was tested with 2 senior fellows. However, the fact that they could complete it in 10 minutes does not indicate that an intern can complete the same task in 10 minutes.
We do not disagree with this fact, although given that our interns and residents are the ones doing rounds and data gathering on patients, it is hard to know whether our interns or fellows would be more facile with this aspect of the system. As stated, the time chosen was based to mimic the daily workflow in our ICU ad the typical amount of time the spent on preparing for rounds on a patient, given their time of entry to the ICU and the number of patients on their census.

The residents were given a signout before examining the medical record. Was this a standardized signout?
Yes,

1. While their finding that the participants detected an average of 41% of the errors is concerning, the investigators’ study design may actually be overly optimistic. For example, the fact that the participants were observed during their data collection and were aware that they would have to present the case later undoubtedly creates a significant Hawthorne effect. The participants may be more vigilant about their data collection because they are aware of being directly observed and tested. In addition, because discussion of EHR use has been integrated into weekly M & M conference, it is likely that the participants, especially the fellows, are more cognizant of the errors that commonly occur with EHR than trainees at other institutions with EHR. While this would bias the results towards improved recognition of errors, this still needs to be discussed further in the discussion precisely because the low rate of error recognition is so alarming. It is important to demonstrate that potential biases from study design would actually bias the results towards the null suggesting that the true error recognition rate may be even lower in actual clinical practice.

This is an excellent point and the discussion has been modified to suggest this.

2.The authors should clarify that the inability of trainees to recognize errors or deteriorating medical condition does not necessarily imply that poor outcomes will result. The errors may be picked up by the pharmacist, the attending or nurse or it may be detected later. Indeed, while “massive amount of information” with data overload is one problem with standard EHRs, error rate and its implied consequences may be exaggerated in this simulation because one person was responsible for finding every error rather than a team of clinicians.

This has been clarified in the discussion section.

Other specific comments about the results, discussion:

Results:
The EHR is just one tool a clinician uses. Some things such as the sedation level of the patient may be easier to assess by a physical exam rather than by searching the EHR. Other things may have been found out through other methods like communication with other health professionals.

The reviewer is correct that the EHR is not the only source of data in the ICU, although in one recent abstract, less than 50% of ICU practioners performed a traditional physical exam in the ICU. It is our hope that future studies will integrate the EHR with either standardized patients or high fidelity simulators, bedside monitors and ventilators to better address each individual components contribution to error recognition.

Discussion:
In citation 22 the subjects were not asked question related to “whether the patient was bleeding”. They were informed that the patient was bleeding and then asked questions like is there adequate IV access, is there a type and screen active.
We apologize for this confusion and this has been corrected

Would it be better to perform simulation training with end-users before purchasing an EHR?
We believe that eventually, these data will demonstrate that simulation training should be used as part of institutional EHR training and optimally be used to help inform the institution’s IT department on the functionality and potential unintended consequences of EHR customization. Further, use of simulation should hopefully be used to inform on the consequences of changes made to the EHR user interface and design at the corporation level as well. It is conceivable that simulation could also have a role in the RFP process of a institutions choice for an EHR, although it would be more difficult to interpret the significance of the results among subjects completely naïve to the system.

Tables:
Some of the items should be detailed and quantified further (increase in WBCs, new fever, recognition of fluid balance)
This has been added in.

Reviewer #2
GOALS
One goal appears to be about using simulation to assess how providers using an EHR can detect evidence of medical errors in the record (p2, Article Focus).
So your outcomes are detection rates, plus a secondary outcome of patterns of navigation in the EHR.

On p 17 line 56, you mention that “the goal of the simulation is to test the system under high-stress/dangerous situations” (and I assume “test” means testing the user/EHR system on the task of

problem detection). You mention this it as a mitigation of a limitation, but I would argue that if your goal is to look at patterns of problem detection, then having a case with lots of problems may be appropriate.
The reviewer is correct about the inference of the word “test” in this context, and yes, while having a case with a number of issues is appropriate given the overall goal, we felt it was important to also present this as a mitigation of a limitation as well.

On p8, the goal is described as development of simulated ICU patient encounter in EHR*, as part of an effort to train better use of EHR so providers can better detect errors and manage data overload. (*I would describe it more as a simulated handoff or rounding task involving a rich multi-day record of a complex case).
The reviewer is correct as that is also a valid way to describe the study. However, given that the goal of this pilot study is to form the foundation for testing interventions aimed at altering the EHR/user interface (screen redesign, or redesign of new user education) without changing the script for signout or handoff, we chose the presented wording. However, the reviewer is correct that this infrastructure can also be easily adapted to evaluate efficacy of modifications to the signout/handoff procedure as well

Your design also includes comparisons across clinical training levels (fellows vs. residents vs. interns).
This is correct and we felt was important to help provide guidance for determining sample size for future interventional studies aimed at modulation of either EHR training or modulation of the use interface.

You also introduce training on p7, and you mention the lack of any training intervention results as a limitation of the study (p2). The study demonstrates an assessment method, rather than a training intervention. If you are going to spend time on the role of simulation in training, I think you should be more clear and explicit about how this study fits in. Using simulators in training, and using simulators for assessment of training (or for assessment of user interface designs) are different things.

The reviewer is correct about the distinction between assessment of training and training itself. The goal of this pilot study was to use simulation to assess EHR use and to establish baseline data so that can use this as a training technique as well. Hence the overlap between assessment and training itself in this case.

The difference between the clinical levels is very interesting, and demonstrates a type of validity (that your protocol can detect differences between these groups). However, you are not explicit about how the comparison across clinical levels fits in with your goals.
We are currently investigating this very issue. The goal of this pilot study was to provide enough data for the larger study we are currently undertaking. The differences between interns and non-interns does not necessarily affect our goals, but rather helps to effectively power our ongoing and future studies.

FYI: One way differences in expertise are used in instructional systems R&D is to study the patterns of use that the more successful practitioners do, and use that to inform the content of training for less experienced people. [See: Crandall B, Klein G, Hoffman R. Working Minds: A Practitioner's Guide to Cognitive Task Analysis (Bradford Books). The MIT Press; 2006.]. What did the fellows do that (presumably) helped them cope with the information overload better?
Related to this: Is there any more you can say about how some people used a larger number of screens in the same 10 minutes? Navigation techniques are one way people cope adapt to

information overload [Watts-Perotti J, Woods D. How Experienced Users Avoid Getting Lost in Large Display Networks. International Journal of Human-Computer Interaction. 1999;11(4):269-299.]

This is a fascinating question and unfortunately we do not have yet enough data to comment on this issue further. We are currently embarking on a more complex usability study with this particular simulation exercise and the results will be the focus of a separate study.

FIDELITY

All simulations represent a sub-set of reality. The question of fidelity is really about which sub-set of reality, and which mode of representation, is relevant to the research question. See:

- Rudolph JW, Simon R, Raemer DB. Which reality matters? Questions on the path to high engagement in healthcare simulation. Simulation in Healthcare. 2007;2(3):161-163.; and
- McCurdy M, Connors C, Pyrzak G et al. Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. ACM; 2006:1233-1242.

Given that your research question is about problem detection via the electronic record, the fact that the test took place in the same physical environment as is normally used may not be especially relevant to your research question. One simulation constraint that may be relevant is that the physician received a written summary of the patient's history, rather than a warm handoff. What questions would an incoming physician ask the outgoing physician about the patient? Would communication with other staff also affect problem detection?

We believe that performing the simulation in the ICU, with the ambient background noises, alarms etc.. is actually quite important for assessing the efficacy of EHR use and better recapitulates the use of the system in clinical life. We agree that the use of a warm handoff and the ability to ask questions would also affect the results of the study and would more recapitulate the real-life workflow. However, we wanted to assure that each subject received the exact amount of information going into the simulation and thus had to find out the same information solely through interaction with the EHR. We believe that by controlling for these variables we can better interpret not only these results, but the results of our ongoing studies modifying the interface itself.

Note that the types of cases (and even the modes of representation) useful for training are not necessarily the same as those for assessment.

ANALYSIS & RESULTS

I'd like to see more on which medical errors are considered serious and which are not. Relatedly, you use categories in Table 1 (changes in patient condition, medication errors, failure to adhere to best practice), but these categories are not used in presentation of results.

This is extremely difficult and subjective based on the definition of the term "serious". For example, some of the errors are immediately life threatening (drop in blood pressure) while others, such as poor glycemic control are associated with increased mortality on a population level (of ICU patients). Further

Can you report anything about the participants' familiarity with EPIC, or EHRs in general?

All of the users had used EPIC throughout their training and all had received institution specific training at the beginning of their training. Further, there are no specific screens used only in the ICU (same workflow as the inpatient environment) so familiarity with the system was not an issue.

I'm not sure if having both figure 2 and figure 3 is necessary.

With figure 2, % correct (instead of N correct vs. N incorrect) would be more consistent with some of

the other figures.
Figure 2 has been corrected

Figure 5 should use the more informative terms used in text.

MISC
Missing word p 5 line 44
Typo p2 Line 34
These have been corrected.

VERSION 2 – REVIEW

REVIEWER	Michelle Ng Gong MD Associate Professor in Medicine and pidemiology and Population Health Director of Critical Care Research Montefiore Medical Center Albert Einstein College of Medicine No competing interest
REVIEW RETURNED	06-Mar-2013

GENERAL COMMENTS	The revised manuscript is overall improved although I still have several comments for the authors to improve the clarity of their reporting. 1) While the authors indicated that all participants received institutional specific training with the EHR in the methods, they still did not clearly indicate what that training is. Training on EHR can differ greatly from institution to institution and range from no training, online tutorial, to dedicated course ranging from an hour to a day. It is important for the readers to understand the extent of the "institutional training" and how it may or may not be contribute to the error rate. Indicating that all participant got institutional training is not adequate to address this reviewer's concern. 2) In their response, the authors indicated that there is no ICU specific interface in their institution's EPIC implementation. This then raises another issue that was not discussed by the authors. Did the lack of customization with ICU specific interfaces and screens contribute to the high error rate? This is especially pertinent as the authors discuss at length why the ICU is particularly vulnerable to errors from EHR and how prior published failures of EHR implementation in the ICU was partly attributed to lack of customization. 3) Table 1 needs clarification. What is meant by EHR category of error? What is meant by structure and time, cognition, customization, and over completeness? How is failure to detect and treat hyperglycemia an error of overcompleteness? 4) In the text, there was a significant difference in error detection between fellows and residents for tidal volume and daily awakening. But in figure 3, only tidal volume was marked as significantly different.

REVIEWER	Michael W. Smith, PhD Health Science Specialist/Human Factors Engineer Houston VA HSR&D Center of Excellence & Baylor College of Medicine USA

	I have no competing interests to declare.
REVIEW RETURNED	08-Mar-2013

GENERAL COMMENTS	This is a good study with useful findings. They have demonstrated the utility of simulations for studying this aspect of EHRs; and have established a method of case development, and have obtained baseline data for further studies. Strengths of this study include: the nature of the simulation content (a complex multi-day case); the means of knowledge elicitation (the familiar and face-valid act of rounding/reporting); and the familiar nature of the UI (esp. with users having their own UI customization settings). These are important because of their relevance to the issue of navigating through the EHR, integrating information dispersed throughout the record, and detecting problems. These are particular aspects of the simulation & assessment exercise, with particular relevance to the behavior being assessed. I believe this paper would be better if it was more precise about its particular strengths, instead of relying so much on the simplification of “high fidelity”. After all, the simulation captures some aspects of the reality of problem detection in the EHR very well, but other aspects (that could be argued to be a component of fidelity) it appropriately sets aside as not relevant for this pilot study. This is not a simulation-specific journal, so a discussion of the nature of simulator fidelity is inappropriate. Rather, there are two sentences I suggest should be revised. They are overly broad statements about simulator fidelity. Given that the paper stresses the fidelity of the simulation as a major contribution, and as a point of contrast with other studies, more precision is appropriate. Pg 7, line 39 “In order for simulation to be effective, however, there must be specific attention given to creating psychological and functional fidelity, i.e. recreating the true “feel” of the goal environment 26.” I suggest something like ‘Full task training via simulation benefits from psychological and functional fidelity’ in order to avoid blurring between simulator effectiveness for training vs. for assessment; and to avoid implying that higher fidelity is always better (Maran & Glavin mention situations where low fidelity is appropriate). Pg 13, line 25 “The growing use of simulation as a tool for assessing competency and improving patient safety has established that both the creation of high-fidelity simulations as well as providing immediate feedback to subjects at the conclusion of the simulation are critical towards achieving maximal benefit from the exercise.” In trying to transition to the list of strengths of the simulation, this sentence is conflating several things. Prompt feedback is good for training. Creating simulations has a lot of potential benefit for assessment and training, which can in turn help patient safety. The growing use of simulation for training and for assessment indicates that it is an important topic. However, the emphasis on high fidelity in the growing use of medical simulation is something debated in the literature.

VERSION 2 – AUTHOR RESPONSE

Reviewer: Michelle Ng Gong MD
1) While the authors indicated that all participants received institutional specific training with the EHR in the methods, they still did not clearly indicate what that training is. Training on EHR can differ greatly from institution to institution and range from no training, online tutorial, to dedicated course ranging from an hour to a day. It is important for the readers to understand the extent of the "institutional training" and how it may or may not be contribute to the error rate. Indicating that all participant got institutional training is not adequate to address this reviewer's concern.

We apologize for the lack of detail. We have added the specifics for what Institutional training entails into the methods section. “. This training was standard for all residents and fellows at the beginning of their training and comprised of 1.5 days of small group instruction with one of the institutions dedicated EHR trainers. Training involved hands on use with the system and included tasks such as data retrieval, data entry and instructions on customization. Users were expected to complete a set number of tasks in each of these areas prior to completion.

2) In their response, the authors indicated that there is no ICU specific interface in their institution's EPIC implementation. This then raises another issue that was not discussed by the authors. Did the lack of customization with ICU specific interfaces and screens contribute to the high error rate? This is especially pertinent as the authors discuss at length why the ICU is particularly vulnerable to errors from EHR and how prior published failures of EHR implementation in the ICU was partly attributed to lack of customization.

We apologize for this confusion. The standard data sheets which are used throughout the inpatient environment were designed initially for the ICU. Therefore, what we meant to imply is that while our data sheets are designed for the ICU, they are no longer ICU specific as they are used universally throughout the inpatient environment and thus not having prior exposure to these screens is not a confounder in this case. We have modified this area of the paper to reflect this.

3) Table 1 needs clarification. What is meant by EHR category of error? What is meant by structure and time, cognition, customization, and over completeness? How is failure to detect and treat hyperglycemia an error of overcompleteness?

These are all different types of ways in which errors commonly used in EHR taxonomy, this was best spelled out in an article by Ash et al in 2004 and this reference has been added into the legend for Table 1. s. Based on the number of places issues related to glycemic control and IV fluids appear within our medical record, we felt that for our specific system, this did represent an overcompleteness issue.

4) In the text, there was a significant difference in error detection between fellows and residents for tidal volume and daily awakening. But in figure 3, only tidal volume was marked as significantly different.

We apologize for the confusion, the symbol was lost at one point during file conversion and we missed this omission. The figure has been corrected.

Reviewer: Michael W. Smith, PhD

This is not a simulation-specific journal, so a discussion of the nature of simulator fidelity is inappropriate. Rather, there are two sentences I suggest should be revised. They are overly broad statements about simulator fidelity. Given that the paper stresses the fidelity of the simulation as a major contribution, and as a point of contrast with other studies, more precision is appropriate.

Pg 7, line 39 “In order for simulation to be effective, however, there must be specific attention given to creating psychological and functional fidelity, i.e. recreating the true “feel” of the goal environment 26.”

I suggest something like ‘Full task training via simulation benefits from psychological and functional fidelity’ in order to avoid blurring between simulator effectiveness for training vs. for assessment; and to avoid implying that higher fidelity is always better (Maran & Glavin mention situations where low fidelity is appropriate).

We have modified this sentence.

Pg 13, line 25 “The growing use of simulation as a tool for assessing competency and improving patient safety has established that both the creation of high-fidelity simulations as well as providing immediate feedback to subjects at the conclusion of the simulation are critical towards achieving maximal benefit from the exercise.”

In trying to transition to the list of strengths of the simulation, this sentence is conflating several things.
Prompt feedback is good for training.
Creating simulations has a lot of potential benefit for assessment and training, which can in turn help patient safety.
The growing use of simulation for training and for assessment indicates that it is an important topic.
However, the emphasis on high fidelity in the growing use of medical simulation is something debated in the literature.

We have modified this sentence.